

Inexact Proximal Gradient Methods for Non-convex and Non-smooth Optimization

Bin Gu

Zhouyuan Huo

Heng Huang

*Department of Computer Science and Engineering
University of Texas at Arlington*

JSGUBIN@GMAIL.COM

ZHOUYUAN.HUO@MAVS.UTA.EDU

HENG@UTA.EDU

Abstract

Non-convex and non-smooth optimization plays an important role in machine learning. Proximal gradient method is one of the most important methods for solving the non-convex and non-smooth problems, where a proximal operator need to be solved exactly for each step. However, in a lot of problems the proximal operator does not have an analytic solution, or is expensive to obtain an exact solution. In this paper, we propose inexact proximal gradient methods (not only a basic inexact proximal gradient method (IPG), but also a Nesterov's accelerated inexact proximal gradient method (AIPG)) for non-convex and non-smooth optimization, which tolerate an error in the calculation of the proximal operator. Theoretical analysis shows that IPG and AIPG have the same convergence rates as in the error-free case, provided that the errors decrease at appropriate rates.

Keywords: Non-convex optimization, non-smooth optimization, proximal gradient, inexact proximal operator, Nesterov's accelerated method

1. Introduction

Non-convex and non-smooth optimization plays an important role in machine learning. In particular, we consider the composite non-convex and non-smooth optimization problems of the form

$$\min_{x \in \mathbb{R}^N} f(x) = g(x) + h(x) \quad (1)$$

where $g : \mathbb{R}^N \mapsto \mathbb{R}$ is smooth and possibly non-convex, $h : \mathbb{R}^N \mapsto \mathbb{R}$ is non-smooth and possibly non-convex. In machine learning, the function $g(x)$ is usually related to the loss term, such as least squares loss (Friedman et al., 2001), Huber loss (Friedman et al., 2001), correntropy induced loss (He et al., 2011), and so on. Note that among these, a lot of loss terms are non-convex. For example, the correntropy induced loss (He et al., 2011; Feng et al., 2015; Chen and Wang, 2016) are non-convex and smooth, which are frequently used for robust regression or classification. The loss term in semi-supervised SVM (Chapelle and Zien, 2005; Chapelle et al., 2006) is non-convex due to the symmetric (sigmoid) loss on the unlabeled samples. The function $h(x)$ is related to the regularization term, such as l_1 -norm (Tibshirani, 1996), Capped- l_1 penalty (Zhang, 2010), nuclear-norm (Hsieh and Olsen, 2014) and so on, which may be convex, also could be non-convex.

Proximal gradient method is one of the most important methods for solving the above mentioned problems. There have been several works for proximal gradient methods. For convex problems, the basic proximal gradient (PG) method (Beck and Teboulle, 2009) has the convergence rate $O(\frac{1}{T})$, and the Nesterov’s accelerated proximal gradient (APG) method has the convergence rate $O(\frac{1}{T^2})$ (Beck and Teboulle, 2009), where T is the number of iterations. For non-convex problems, Ghadimi and Lan (2016) considered that only $g(x)$ is non-convex, and prove that the convergence rate of APG method. Boş et al. (2016) considered that both of $g(x)$ and $h(x)$ are non-convex, and proved the convergence rate of PG method. Li and Lin (2015) considered that both of $g(x)$ and $h(x)$ are non-convex, and proved the convergence rate of APG method. In addition to the above batch proximal gradient methods, there are also stochastic and online proximal gradient methods (Duchi and Singer, 2009; Xiao and Zhang, 2014) which are not the focus of this paper.

The key of the proximal gradient method is to solve a proximal operator exactly for each step, as following

$$\text{Prox}_t(y) = \arg \min_{x \in \mathbb{R}^N} \frac{1}{2\gamma} \|x - y\|^2 + h(x) \quad (2)$$

where γ is the step size and determined by line search. However, the proximal operator does not have an analytic solution, or is expensive to obtain the solution exactly, in a lot of problems such as the overlapped group Lasso (Jacob et al., 2009; Jenatton et al., 2010), nuclear-norm minimization (Hsieh and Olsen, 2014), OSCAR (Zhong and Kwok, 2012) and so on. In order to handle the complex proximal operators in the above problems, Schmidt et al. (2011) proposed inexact proximal gradient (IPG) and inexact Nesterov’s accelerated proximal gradient (IAPG) methods for the convex problems, and proved the convergence rates. Later, Villa et al. (2013) proposed an accelerated and inexact forward-backward splitting method (AIFB, actually AIFB is IAPG) and proved the convergence rates. As mentioned previously, non-convex and non-smooth optimization is important in machine learning. To the best of our knowledge, extending IPG and IAPG methods to non-convex and non-smooth problems and providing the corresponding convergence analysis is still an open problem.

Table 1: Representative (exact and inexact) proximal gradient algorithms. (C and NC are the abbreviations of convex and non-convex respectively.)

Algorithm	Proximal	Accelerated	$g(x)$	$h(x)$	Reference
PG+APG	Exact	Yes	C	C	Beck and Teboulle (2009)
APG	Exact	Yes	C+NC	C	Ghadimi and Lan (2016)
PG	Exact	No	NC	NC	Boş et al. (2016)
APG	Exact	Yes	C+NC	C+NC	Li and Lin (2015)
IPG+IAPG	Inexact	Yes	C	C	Schmidt et al. (2011)
AIFB	Inexact	Yes	C	C	Villa et al. (2013)
IPG+IAPG	Inexact	Yes	C+NC	C+NC	Ours

To address the open problem as mentioned above, in this paper, we extend IPG and IAPG methods to non-convex and non-smooth problems, which tolerate an error in the

calculation of the proximal operator. We also provide the theoretical analysis to show that IPG and AIPG have the same convergence rates as in the error-free case, provided that the errors decrease at appropriate rates.

We organize the rest of the paper as follows. In Section 2, we give some preliminaries. In section 3, we propose our IPG and AIPG algorithms. In Section 4, we prove the convergence rates for IPG and AIPG for the non-convex and non-smooth problems. Finally, we give some concluding remarks in Section 5.

2. Preliminaries

In this section, we introduce the Lipschitz smooth, ε -approximate subdifferential and ε -approximate Kurdyka-Łojasiewicz (KL) property, which are critical to the convergence analysis of IPG and IAPG methods for non-convex and non-smooth problems.

Lipschitz smooth: For the smooth functions $g(x)$, we have the Lipschitz constant L for $\nabla g(x)$ as following.

Assumption 1 L is the Lipschitz constant for $\nabla g(x)$. Thus, $\forall x$ and $\forall y$, L -Lipschitz smooth can be presented as

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\| \quad (3)$$

Equivalently, L -Lipschitz smooth can also be written as the formulation (4).

$$g(x) \leq g(y) + \langle \nabla g(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 \quad (4)$$

ε -approximate subdifferential: Because inexact proximal gradient is used in this paper, an ε -approximate proximal operator may produce an ε -approximate subdifferential. In the following, we give the definition of ε -approximate subdifferential (Bertsekas et al., 2003) which will be used in the analysis for the case that $h(x)$ is convex.

Definition 1 Given a convex function $h(x) : \mathbb{R}^N \mapsto \mathbb{R}$ and a positive scalar ε , the ε -approximate subdifferential of $h(x)$ at a point $x \in \mathbb{R}^N$ (denoted as $\partial_\varepsilon h(x)$) is

$$\partial_\varepsilon h(x) = \{d \in \mathbb{R}^N : h(y) \geq h(x) + \langle d, y - x \rangle - \varepsilon\} \quad (5)$$

Based on Definition 1, if $d \in \partial_\varepsilon h(x)$, we say that d is an ε -approximate subgradient of $h(x)$ at a point x .

ε -approximate KL property: Originally, KL property is introduced to analysis the convergence rate of exact proximal gradient methods for non-convex and non-smooth problems (Li and Lin, 2015; Boţ et al., 2016). Because this paper focuses on the inexact proximal gradient methods, correspondingly we give the ε -approximate KL property as following. In Definition 2, the function $\text{dis}(x, S) = \min_{y \in S} \|x - y\|$, where S is a subset of \mathbb{R}^N .

Definition 2 A function $f(x) = g(x) + h(x) : \mathbb{R}^N \rightarrow (-\infty, +\infty]$ is said to have the ε -KL property at $\bar{u} \in \{x \in \mathbb{R}^N : \nabla g(u) + \partial_\varepsilon h(u) \neq \emptyset\}$, if there exists $\eta \in (0, +\infty]$, a neighborhood U of \bar{u} and a function $\varphi \in \Phi_\eta$, such that for all $u \in U \cap \{u \in \mathbb{R}^N : f(\bar{u}) < f(u) < f(\bar{u}) + \eta\}$, the following inequality holds

$$\varphi'(f(u) - f(\bar{u}))\text{dist}(\mathbf{0}, \nabla g(u) + \partial_\varepsilon h(u)) \geq 1 \quad (6)$$

where Φ_η stands for a class of functions $\varphi : [0, \eta) \rightarrow \mathbb{R}^+$ satisfying: (1) φ is concave and continuously differentiable function on $(0, \eta)$; (2) φ is continuous at 0, $\varphi(0) = 0$; and (3) $\varphi'(x) > 0, \forall x \in (0, \eta)$.

3. Algorithms

To make the paper self-contained, we present IPG (i.e., Algorithm 1) and IAPG (i.e., Algorithm 2). Actually, IPG is exactly same with the one in (Schmidt et al., 2011). To provide the convergence rate of IAPG for the non-convex case, we give a different version of IAPG with the one used in (Schmidt et al., 2011). And, the framework of IAPG is used in (Li and Lin, 2015) for the non-convex optimization, in which we replace the exact proximal operator with an inexact proximal operator.

3.1 Basic inexact proximal gradient method

We first present the basic inexact proximal gradient method (i.e. IPG). The framework of IPG is same with the one of PG. As mentioned previously, the key of PG is to compute an exact proximal operator (2) in each step. Correspondingly, the key of IPG is to compute an inexact proximal operator for each step as following

$$x \in \text{Prox}_{\gamma g}^\varepsilon(y) = \left\{ z \in \mathbb{R}^N : \frac{1}{2\gamma} \|z - y\|^2 + h(z) \leq \varepsilon + \min_x \frac{1}{2\gamma} \|x - y\|^2 + h(x) \right\} \quad (7)$$

where ε denote an error in the proximal operator. As discussed in (Tappenden et al., 2016), there are several methods to compute the inexact proximal operator, in which the most popular method is using a primal dual algorithm to control the dual gap.

Algorithm 1 Basic inexact proximal gradient method (IPG)

Input: m, ε_k ($k = 1, \dots, m$), $\gamma < \frac{1}{L}$.

Output: x_m .

- 1: Initialize $x_0 \in \mathbb{R}^d$.
 - 2: **for** $k = 1, \dots, m$ **do**
 - 3: Compute $x_k \in \text{Prox}_{\gamma g}^{\varepsilon_k}(x_{k-1} - \gamma \nabla g(x_{k-1}))$.
 - 4: **end for**
-

3.2 Accelerated inexact proximal gradient method

As we known, APG and IAPG uses a momentum term to accelerate the proximal gradient method for the convex optimization. However, as mentioned in (Li and Lin, 2015), traditional Nesterov's accelerated method may encounter a bad momentum term for the non-convex optimization. To address the bad momentum term, Li and Lin (2015) added a proximal operator as a monitor to make the objective function sufficient descent. To make our IAPG works well for the non-convex optimization, our IAPG follows the framework of APG used in (Li and Lin, 2015). Thus, we compute two inexact proximal operators $z_{k+1} \in \text{Prox}_{\gamma g}^{\varepsilon_k}(y_k - \gamma \nabla g(y_k))$ and $v_{k+1} \in \text{Prox}_{\gamma g}^{\varepsilon_k}(x_k - \gamma \nabla g(x_k))$, where v_{k+1} is a monitor to make the objective function sufficient descent. Specifically, our IAPG is presented in Algorithm 2.

Algorithm 2 Accelerated inexact proximal gradient method (AIPG)

Input: m, ε_k ($k = 1, \dots, m$), $t_0 = 0, t_1 = 1, \gamma < \frac{1}{L}$.

Output: x_{m+1} .

- 1: Initialize $x_0 \in \mathbb{R}^d$, and $x_1 = z_1 = x_0$.
 - 2: **for** $k = 1, 2, \dots, m$ **do**
 - 3: $y_k = x_k + \frac{t_{k-1}}{t_k}(z_k - x_k) + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$.
 - 4: Compute z_{k+1} such that $z_{k+1} \in \text{Prox}_{\gamma g}^{\varepsilon_k}(y_k - \gamma \nabla g(y_k))$.
 - 5: Compute v_{k+1} such that $v_{k+1} \in \text{Prox}_{\gamma g}^{\varepsilon_k}(x_k - \gamma \nabla g(x_k))$.
 - 6: $t_{k+1} = \frac{\sqrt{4t_k^2+1}+1}{2}$.
 - 7: $x_{k+1} = \begin{cases} z_{k+1} & \text{if } f(z_{k+1}) \leq f(v_{k+1}) \\ v_{k+1} & \text{otherwise} \end{cases}$
 - 8: **end for**
-

4. Convergence Analysis

In this section, we prove the convergence rates of our IPG and AIPG for the non-convex optimization. Specifically, we first prove that IPG and AIPG converge to a critical point in the convex and non-convex optimization (Theorem 3) if $\sum_{k=1}^m \varepsilon_k < \infty$. Next, we then prove that IPG has the convergence rate $O(\frac{1}{T})$ for the non-convex optimization (Theorem 4) when the errors decrease at an appropriate rate. Then, we prove that the convergence rates for AIPG in the non-convex optimization (Theorem 7). In addition, because our AIPG is different to the one in (Schmidt et al., 2011), we also prove the convergence rate $O(\frac{1}{T^2})$ for the convex optimization (Theorem 8) when the errors decrease at an appropriate rate.

We first prove that IPG and AIPG converge to a critical point for the convex or non-convex optimization (Theorem 3) if $\sum_{k=1}^m \varepsilon_k < \infty$.

Theorem 3 *If $\{\varepsilon_k\}$ is a decreasing sequence and $\sum_{k=1}^m \varepsilon_k < \infty$, we have $\mathbf{0} \in \lim_{k \rightarrow \infty} \nabla g(x_k) + \partial_{\varepsilon_k} h(x_k)$ for IPG and AIPG in the convex and non-convex optimizations.*

Proof We prove that $\mathbf{0} \in \lim_{k \rightarrow \infty} \nabla g(x_k) + \partial_{\varepsilon_k} h(x_k)$ for AIPG in the convex and non-convex optimizations if $\sum_{k=1}^m \varepsilon_k < \infty$. The proof for IPG can be provided similarly.

According to line 5 in Algorithm 1 and (7), we have that

$$\langle \nabla g(x_k), v_{k+1} - x_k \rangle + \frac{1}{2\gamma} \|v_{k+1} - x_k\|^2 + h(v_{k+1}) \leq h(x_k) + \varepsilon_k \quad (8)$$

Thus, we have that

$$\begin{aligned} f(v_{k+1}) &= g(v_{k+1}) + h(v_{k+1}) \\ &\leq g(x_k) + \langle \nabla g(x_k), v_{k+1} - x_k \rangle + \frac{L}{2} \|v_{k+1} - x_k\|^2 \\ &\quad + h(x_k) - \langle \nabla g(x_k), v_{k+1} - x_k \rangle - \frac{1}{2\gamma} \|v_{k+1} - x_k\|^2 + \varepsilon_k \\ &= f(x_k) - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|v_{k+1} - x_k\|^2 + \varepsilon_k \end{aligned} \quad (9)$$

If $f(z_{k+1}) \leq f(v_{k+1})$, we have $x_{k+1} = z_{k+1}$ and $f(x_{k+1}) = f(z_{k+1}) \leq f(v_{k+1})$. If $f(z_{k+1}) > f(v_{k+1})$, we have $x_{k+1} = v_{k+1}$ and $f(x_{k+1}) = f(v_{k+1})$. Thus, we have that

$$f(x_{k+1}) \leq f(x_k) - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|v_{k+1} - x_k\|^2 + \varepsilon_k \quad (10)$$

By summing the the inequality (10) over $k = 1, \dots, m$, we obtain

$$f(x_{m+1}) \leq f(x_0) - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \sum_{k=1}^m \|v_{k+1} - x_k\|^2 + \sum_{k=1}^m \varepsilon_k \quad (11)$$

Same with the analysis for (27) in Theorem 23, we have that

$$\sum_{k=1}^m \|v_{k+1} - x_k\|^2 \leq \frac{1}{\frac{1}{2\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \frac{1}{\frac{1}{2\gamma} - \frac{L}{2}} \sum_{k=1}^m \varepsilon_k \quad (12)$$

We assume that $\sum_{k=1}^m \varepsilon_k < \infty$. Thus, $\sum_{k=1}^m \|v_{k+1} - x_k\|^2 \leq \frac{1}{\frac{1}{2\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \frac{1}{\frac{1}{2\gamma} - \frac{L}{2}} \sum_{k=1}^m \varepsilon_k < \infty$. So we have that

$$\lim_{k \rightarrow \infty} \|v_{k+1} - x_k\|^2 = 0 \quad (13)$$

In addition, we have $\lim_{k \rightarrow \infty} \varepsilon_k = 0$. Because v_{k+1} is a ε_k -optimal solution to the proximal problem (7), according to Lemma 2 in Schmidt et al. (2011), there exists f_k such that $\|f_k\| \leq \sqrt{2\gamma\varepsilon_k}$ and

$$\begin{aligned} 0 &\in \frac{1}{\gamma} (x_k - v_{k+1} - \gamma \nabla g(x_k) - f_k) - \partial_{\varepsilon_k} h(v_{k+1}) \\ &= \frac{1}{\gamma} (x_k - v_{k+1} - f_k) - \nabla g(x_k) + \nabla g(v_{k+1}) - \nabla g(v_{k+1}) - \partial_{\varepsilon_k} h(v_{k+1}) \end{aligned} \quad (14)$$

Thus, we have

$$\frac{1}{\gamma} (x_k - v_{k+1} - f_k) - \nabla g(x_k) + \nabla g(v_{k+1}) \in \nabla g(v_{k+1}) + \partial_{\varepsilon_k} h(v_{k+1}) \quad (15)$$

In addition, we have

$$\left\| \frac{1}{\gamma} (x_k - v_{k+1} - f_k) - \nabla g(x_k) + \nabla g(v_{k+1}) \right\| \leq \left(\frac{1}{\gamma} + L \right) \|x_k - v_{k+1}\| + \sqrt{\frac{2\varepsilon_k}{\gamma}} \quad (16)$$

Thus, we have

$$\begin{aligned} &\lim_{k \rightarrow \infty} \left\| \frac{1}{\gamma} (x_k - v_{k+1} - f_k) - \nabla g(x_k) + \nabla g(v_{k+1}) \right\| \\ &\leq \lim_{k \rightarrow \infty} \left(\left(\frac{1}{\gamma} + L \right) \|x_k - v_{k+1}\| + \sqrt{\frac{2\varepsilon_k}{\gamma}} \right) = 0 \end{aligned} \quad (17)$$

Based on (15) and (17), we have that

$$\mathbf{0} \in \lim_{k \rightarrow \infty} \nabla g(v_k) + \partial_{\varepsilon_k} h(v_k) \quad (18)$$

Because $\lim_{k \rightarrow \infty} \|v_{k+1} - x_k\|^2 = 0$ as proved in (13), we have that

$$\mathbf{0} \in \lim_{k \rightarrow \infty} \nabla g(x_k) + \partial_{\varepsilon_k} h(x_k) \quad (19)$$

This completes the proof. ■

4.1 IPG for nonconvex optimization

Based on (15), we similarly have

$$\frac{1}{\gamma} (x_{k-1} - x_k - f_k) - \nabla g(x_{k-1}) + \nabla g(x_k) \in \nabla g(x_k) + \partial_{\varepsilon_k} h(x_k) \quad (20)$$

for IPG. Further, similar with (16), we have

$$\left\| \frac{1}{\gamma} (x_{k-1} - x_k - f_k) - \nabla g(x_{k-1}) + \nabla g(x_k) \right\| \leq \left(\frac{1}{\gamma} + L \right) \|x_k - x_{k-1}\| + \sqrt{\frac{2\varepsilon_k}{\gamma}} \quad (21)$$

Thus, we have that

$$\frac{1}{m} \sum_{k=1}^m \min_{d_k \in \partial_{\varepsilon_k} h(x_k)} \|\nabla g(x_k) + d_k\|^2 \leq \frac{1}{m} \sum_{k=1}^m \left(\left(\frac{1}{\gamma} + L \right) \|x_k - x_{k-1}\| + \sqrt{\frac{2\varepsilon_k}{\gamma}} \right) \quad (22)$$

Based on (22), we use $\frac{1}{m} \sum_{k=1}^m \|x_k - x_{k-1}\|^2$ to analyze the convergence rate in the non-convex setting.

Theorem 4 *For $g(x)$ is nonconvex, and $h(x)$ is convex or nonconvex, we have the following results for IPG:*

1. *If $h(x)$ is convex, we have that*

$$\frac{1}{m} \sum_{k=1}^m \|x_k - x_{k-1}\|^2 \leq \frac{1}{m} \left(2A_m + \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*))} + \sqrt{B_m} \right)^2 \quad (23)$$

where $A_m = \frac{1}{2} \sum_{k=1}^m \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sqrt{\frac{2\varepsilon_k}{\gamma}}$ and $B_m = \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sum_{k=1}^m \varepsilon_k$.

2. *If $h(x)$ is non-convex, we have that*

$$\frac{1}{m} \sum_{k=1}^m \|x_k - x_{k-1}\|^2 \leq \frac{1}{m \left(\frac{1}{2\gamma} - \frac{L}{2} \right)} \left(f(x_0) - f(x^*) + \sum_{k=1}^m \varepsilon_k \right) \quad (24)$$

Proof We first give the proof for the case that $h(x)$ is convex. Since $x_k \in \text{Prox}_{\gamma g}^{\varepsilon_k}(x_{k-1} - \gamma \nabla g(x_{k-1}))$, according to Lemma 2 in (Schmidt et al., 2011), there exists f_k such that $\|f_k\| \leq \sqrt{2\gamma\varepsilon_k}$ and

$$\frac{1}{\gamma}(x_{k-1} - x_k - \gamma \nabla g(x_{k-1}) - f_k) \in \partial_{\varepsilon_k} h(x_k) \quad (25)$$

We have that

$$\begin{aligned} f(x_k) &= g(x_k) + h(x_k) \\ &\leq g(x_{k-1}) + \langle \nabla g(x_{k-1}), x_k - x_{k-1} \rangle + \frac{L}{2} \|x_k - x_{k-1}\|^2 \\ &\quad + h(x_{k-1}) - \left\langle \nabla g(x_{k-1}) + \frac{1}{\gamma}(x_k - x_{k-1} + f_k), x_k - x_{k-1} \right\rangle + \varepsilon_k \\ &= f(x_{k-1}) - \frac{1}{\gamma} \|x_k - x_{k-1}\|^2 + \frac{L}{2} \|x_k - x_{k-1}\|^2 - \left\langle \frac{1}{\gamma} f_k, x_k - x_{k-1} \right\rangle + \varepsilon_k \\ &\leq f(x_{k-1}) - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \|x_k - x_{k-1}\|^2 + \sqrt{\frac{2\varepsilon_k}{\gamma}} \|x_k - x_{k-1}\| + \varepsilon_k \end{aligned} \quad (26)$$

where the first inequality uses (4), the convexity of h and (25). By summing the inequality (26) over $k = 1, \dots, m$, we obtain

$$f(x_m) \leq f(x_0) - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \sum_{k=1}^m \|x_k - x_{k-1}\|^2 + \sum_{k=1}^m \sqrt{\frac{2\varepsilon_k}{\gamma}} \|x_k - x_{k-1}\| + \sum_{k=1}^m \varepsilon_k \quad (27)$$

According to (27), we have that

$$\|x_m - x_{m-1}\|^2 \leq \underbrace{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \left(f(x_0) - f(x_m) + \sum_{k=1}^m \varepsilon_k \right)}_A + \sum_{k=1}^m \underbrace{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sqrt{\frac{2\varepsilon_k}{\gamma}}}_{\lambda_k} \|x_k - x_{k-1}\| \quad (28)$$

According to Lemma 1 in (Schmidt et al., 2011), we have that

$$\begin{aligned} &\|x_m - x_{m-1}\| \\ &\leq \frac{1}{2} \sum_{k=1}^m \lambda_k + \left(A + \left(\frac{1}{2} \sum_{k=1}^m \lambda_k \right)^2 \right)^{\frac{1}{2}} \\ &= \underbrace{\frac{1}{2} \sum_{k=1}^m \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sqrt{\frac{2\varepsilon_k}{\gamma}}}_{A_m} + \left(\underbrace{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x_m)) + \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sum_{k=1}^m \varepsilon_k}_{B_m} + \underbrace{\left(\frac{1}{2} \sum_{k=1}^m \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sqrt{\frac{2\varepsilon_k}{\gamma}} \right)^2}_{A_m} \right)^{\frac{1}{2}} \\ &\leq A_m + \left(\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + B_m + A_m^2 \right)^{\frac{1}{2}} \end{aligned} \quad (29)$$

Because A_k and B_k are increasing sequences, $\forall k \leq m$, we have that

$$\begin{aligned}
& \|x_k - x_{k-1}\| \\
& \leq A_m + \left(\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + B_m + A_m^2 \right)^{\frac{1}{2}} \\
& \leq A_m + \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \sqrt{B_m} + A_m} \\
& \leq 2A_m + \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \sqrt{B_m}}
\end{aligned} \tag{30}$$

According to (27) and (30), we have that

$$\begin{aligned}
& \sum_{k=1}^m \|x_k - x_{k-1}\|^2 \\
& \leq \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x_m)) + \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sum_{k=1}^m \varepsilon_k + \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} \sum_{k=1}^m \sqrt{2L\varepsilon_k} \|x_k - x_{k-1}\| \\
& \leq \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + B_m + 2A_m \left(2A_m + \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \sqrt{B_m}} \right) \\
& \leq \frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + 2\sqrt{B_m} \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + B_m} \\
& \quad + 2A_m \left(2A_m + \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \sqrt{B_m}} \right) \\
& = \left(2A_m + \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \sqrt{B_m}} \right)^2
\end{aligned} \tag{31}$$

Based on (31), we have that

$$\frac{1}{m} \sum_{k=1}^m \|x_k - x_{k-1}\|^2 \leq \frac{1}{m} \left(2A_m + \sqrt{\frac{1}{\frac{1}{\gamma} - \frac{L}{2}} (f(x_0) - f(x^*)) + \sqrt{B_m}} \right)^2 \tag{32}$$

This completes the conclusion for the case that $h(x)$ is convex.

Next, we give the the proof for the case that $h(x)$ is non-convex. According to line 3 in Algorithm 1 and (7), we have that

$$\langle \nabla g(x_{k-1}), x_k - x_{k-1} \rangle + \frac{1}{2\gamma} \|x_k - x_{k-1}\|^2 + h(x_k) \leq h(x_{k-1}) + \varepsilon_k \tag{33}$$

Thus, we have that

$$f(x_k) = g(x_k) + h(x_k) \tag{34}$$

$$\begin{aligned}
&\leq g(x_{k-1}) + \langle \nabla g(x_{k-1}), x_k - x_{k-1} \rangle + \frac{L}{2} \|x_k - x_{k-1}\|^2 \\
&\quad + h(x_{k-1}) - \langle \nabla g(x_{k-1}), x_k - x_{k-1} \rangle - \frac{1}{2\gamma} \|x_k - x_{k-1}\|^2 + \varepsilon_k \\
&= f(x_{k-1}) - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|x_k - x_{k-1}\|^2 + \varepsilon_k
\end{aligned}$$

By summing the inequality (34) over $k = 1, \dots, m$, we obtain

$$f(x_m) \leq f(x_0) - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \sum_{k=1}^m \|x_k - x_{k-1}\|^2 + \sum_{k=1}^m \varepsilon_k \quad (35)$$

Based on (35), we have that

$$\frac{1}{m} \sum_{k=1}^m \|x_k - x_{k-1}\|^2 \leq \frac{1}{m \left(\frac{1}{2\gamma} - \frac{L}{2} \right)} \left(f(x_0) - f(x^*) + \sum_{k=1}^m \varepsilon_k \right) \quad (36)$$

This completes the proof. ■

Remark 5 *Theorem 4 implies that IPG has the convergence rate $O(\frac{1}{T})$ for the non-convex optimization without errors. If $\{\sqrt{\varepsilon_k}\}$ is summable and $h(x)$ is convex, we can also have that IPG has the convergence rate $O(\frac{1}{T})$ for the non-convex optimization. If $\{\varepsilon_k\}$ is summable and $h(x)$ is non-convex, we can also have that IPG has the convergence rate $O(\frac{1}{T})$ for the non-convex optimization.*

4.2 AIPG

In this section, we prove that the convergence rates for AIPG in the non-convex optimization (Theorem 7). In addition, we prove the convergence rate $O(\frac{1}{T^2})$ for the convex optimization (Theorem 8) when the errors decrease at an appropriate rate.

4.2.1 NONCONVEX OPTIMIZATION

To prove the convergence rate of AIPG for nonconvex optimization, we first give Lemma 6. Lemma 6 is an ε approximate version of uniformized KL property which can be proved similarly with the analysis of Lemma 6 in (Bolte et al., 2014). Based on Lemma 6, we prove the convergence rate of AIPG for non-convex optimization (Theorem 7).

Lemma 6 *Let Ω be a compact set and let $f(x) : \mathbb{R}^N \rightarrow (-\infty, +\infty]$ be a proper and lower semicontinuous function. Assume that $f(x)$ is constant on Ω and satisfies the ε -KL property at each point of Ω . Then there exists $\epsilon > 0$, $\eta > 0$ and $\varphi \in \Phi_\eta$, such that for all $\bar{u} \in \Omega$ and all u in the following intersection*

$$\{u \in \mathbb{R}^N : \text{dist}(u, \Omega) < \epsilon\} \cap \{u \in \mathbb{R}^N : f(\bar{u}) < f(u) < f(\bar{u}) + \eta\} \quad (37)$$

the following inequality holds

$$\varphi'(f(u) - f(\bar{u})) \text{dist}(\mathbf{0}, \nabla g(u) + \partial_\varepsilon h(u)) \geq 1 \quad (38)$$

Theorem 7 Assume that g is a nonconvex function with Lipschitz continuous gradients, h is a proper and lower semicontinuous function. If the function f satisfies the ε -KL property, $\varepsilon_k = \alpha \|v_{k+1} - x_k\|^2$, $\alpha \geq 0$, $\frac{1}{2\gamma} - \frac{L}{2} - \alpha \geq 0$ and the desingularising function has the form $\varphi(t) = \frac{C}{\theta} t^\theta$ for some $C > 0$, $\theta \in (0, 1]$, then

1. If $\theta = 1$, there exists k_1 such that $f(x_k) = f^*$ for all $k > k_1$ and AIPG terminates in a finite number of steps, where $\lim_{k \rightarrow \infty} f(x_k) = f^*$.
2. If $\theta \in [\frac{1}{2}, 1)$, there exists k_2 such that for all $k > k_2$

$$f(x_k) - \lim_{k \rightarrow \infty} f(x_k) \leq \left(\frac{d_1 C^2}{1 + d_1 C^2} \right)^{k-k_2} (f(v_k) - f^*) \quad (39)$$

$$\text{where } d_1 = \frac{\left(\frac{1}{\gamma} + L + \sqrt{\frac{2\alpha}{\gamma}} \right)^2}{\frac{1}{2\gamma} - \frac{L}{2} - \alpha}.$$

3. If $\theta \in (0, \frac{1}{2})$, there exists k_3 such that for all $k > k_3$

$$f(x_k) - \lim_{k \rightarrow \infty} f(x_k) \leq \left(\frac{C}{(k - k_3) d_2 (1 - 2\theta)} \right)^{\frac{1}{1-2\theta}} \quad (40)$$

$$\text{where } d_2 = \min \left\{ \frac{1}{2d_1 C}, \frac{C}{1-2\theta} \left(2^{\frac{2\theta-1}{2\theta-2}} - 1 \right) (f(v_0) - f^*)^{2\theta-1} \right\}.$$

Proof We first give the upper bounds for $\|v_{k+1} - x_k\|^2$ and $\text{dist}(\mathbf{0}, \nabla g(v_{k+1}) + \partial_{\varepsilon_k} h(v_{k+1}))$ respectively. From (26) and $f(x_k) \leq f(v_k)$, we have that

$$\begin{aligned} f(v_{k+1}) &\leq f(v_k) - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|v_{k+1} - x_k\|^2 + \varepsilon_k \\ &= f(v_k) - \left(\frac{1}{2\gamma} - \frac{L}{2} - \alpha \right) \|v_{k+1} - x_k\|^2 \end{aligned} \quad (41)$$

Thus, we have that $f(v_{k+1}) \leq f(v_k)$ and

$$\|v_{k+1} - x_k\|^2 \leq \frac{f(v_k) - f(v_{k+1})}{\frac{1}{2\gamma} - \frac{L}{2} - \alpha} \quad (42)$$

From (15), we have that

$$\begin{aligned} \text{dist}(\mathbf{0}, \nabla g(v_{k+1}) + \partial_{\varepsilon_k} h(v_{k+1})) &\leq \left(\frac{1}{\gamma} + L \right) \|x_k - v_{k+1}\| + \sqrt{\frac{2\varepsilon_k}{\gamma}} \\ &= \left(\frac{1}{\gamma} + L + \sqrt{\frac{2\alpha}{\gamma}} \right) \|x_k - v_{k+1}\| \end{aligned} \quad (43)$$

According to (13), we known $\{x_k\}$ and $\{v_k\}$ convergence to the same points. Let Ω be the set that contains all the convergence points of $\{x_k\}$ (also $\{v_k\}$). Because $f(v_{k+1}) \leq f(v_k)$, $\{f(v_k)\}$ is a monotonically decreasing sequence. Thus, $f(v_k)$ has the same value at all the convergence points in Ω , which is denoted as f^* .

Because $\{f(v_k)\}$ is a monotonically decreasing sequence, there exists \widehat{k}_1 such that $f(v_k) \leq f^* + \eta$, $\forall k > \widehat{k}_1$. On the other hand, because $\lim_{k \rightarrow \infty} \text{dis}(v_k, \Omega) = 0$, there exists \widehat{k}_2 such that $\text{dis}(v_k, \Omega) < \epsilon$, $\forall k > \widehat{k}_2$. Let $k > k_0 = \max\{\widehat{k}_1, \widehat{k}_2\}$, we have

$$v_k \in \{v : \text{dis}(v, \Omega) < \epsilon\} \cap \{v : f^* < f(v) < f^* + \eta\} \quad (44)$$

From Lemma 6, there exists a concave function φ such that

$$\varphi'(f(v_k) - f^*) \text{dist}(\mathbf{0}, \nabla g(v_k) + \partial_\varepsilon h(v_k)) \geq 1 \quad (45)$$

Define $r_k = f(v_k) - f^*$. According to (42), (43) and (45), we have that

$$\begin{aligned} 1 &\leq (\varphi'(f(v_k) - f^*) \text{dist}(\mathbf{0}, \nabla g(u) + \partial_\varepsilon h(v_k)))^2 \\ &\leq (\varphi'(r_k))^2 \left(\frac{1}{\gamma} + L + \sqrt{\frac{2\alpha}{\gamma}} \right)^2 \|x_{k-1} - v_k\|^2 \\ &\leq (\varphi'(r_k))^2 \left(\frac{1}{\gamma} + L + \sqrt{\frac{2\alpha}{\gamma}} \right)^2 \frac{f(v_{k-1}) - f(v_k)}{\frac{1}{2\gamma} - \frac{L}{2} - \alpha} \\ &= d_1 (\varphi'(r_k))^2 (r_{k-1} - r_k) \end{aligned} \quad (46)$$

for all $k > k_0$, where $d_1 = \frac{\left(\frac{1}{\gamma} + L + \sqrt{\frac{2\alpha}{\gamma}}\right)^2}{\frac{1}{2\gamma} - \frac{L}{2} - \alpha}$. Because φ has the form of $\varphi(t) = \frac{C}{\theta} t^\theta$, we have $\varphi'(t) = C t^{\theta-1}$. Thus, according to (46), we have that

$$1 \leq d_1 C^2 t^{2\theta-2} (r_{k-1} - r_k) \quad (47)$$

Next, we consider the three cases, i.e., $\theta = 1$, $\theta \in [\frac{1}{2}, 1)$ and $\theta \in (0, \frac{1}{2})$, which are also considered in (Li and Lin, 2015). Same with the analysis of Theorem 3 in (Li and Lin, 2015), we can have the conclusions in Theorem 7. This completes the proof. \blacksquare

4.2.2 CONVEX OPTIMIZATION

In this section, we prove the convergence rate $O(\frac{1}{T^2})$ for the convex optimization (Theorem 8) when the errors decrease at an appropriate rate.

Theorem 8 *Assume that f is convex. For AIPG, we have that*

$$f(x_{k+1}) - f(x^*) \leq \frac{2L}{(m+1)^2} \left(\|x_0 - x^*\| + 2A_m + \sqrt{B_m} \right)^2 \quad (48)$$

where $A_m = \frac{1}{2} \sum_{k=1}^m 2\gamma t_k \sqrt{2\frac{1}{\gamma} \varepsilon_k}$, $B_m = 2\gamma \sum_{k=1}^m t_k^2 \varepsilon_k$.

Proof We have that

$$\begin{aligned} f(z_{k+1}) &= g(z_{k+1}) + h(z_{k+1}) \\ &\leq g(y_k) + \langle \nabla g(y_k), z_{k+1} - y_k \rangle + \frac{L}{2} \|z_{k+1} - y_k\|^2 + h(z_{k+1}) \end{aligned} \quad (49)$$

$$\begin{aligned}
&= g(y_k) + \langle \nabla g(y_k), x - y_k \rangle + \langle \nabla g(y_k), z_{k+1} - x \rangle + \frac{L}{2} \|z_{k+1} - y_k\|^2 + h(z_{k+1}) \\
&\leq g(x) + \langle \nabla g(y_k), z_{k+1} - x \rangle + \frac{L}{2} \|z_{k+1} - y_k\|^2 + h(z_{k+1}) \\
&\leq g(x) + \langle \nabla g(y_k), z_{k+1} - x \rangle + \frac{L}{2} \|z_{k+1} - y_k\|^2 \\
&\quad + h(x) - \left\langle \nabla g(y_k) + \frac{1}{\gamma}(z_{k+1} - y_k + f_k), z_{k+1} - x \right\rangle + \varepsilon_k \\
&= f(x) + \frac{L}{2} \|z_{k+1} - y_k\|^2 - \left\langle \frac{1}{\gamma}(z_{k+1} - y_k), z_{k+1} - y_k + y_k - x \right\rangle + \left\langle \frac{1}{\gamma}f_k, x - z_{k+1} \right\rangle + \varepsilon_k \\
&= f(x) - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \|z_{k+1} - y_k\|^2 - \left\langle \frac{1}{\gamma}(z_{k+1} - y_k), y_k - x \right\rangle + \left\langle \frac{1}{\gamma}f_k, x - z_{k+1} \right\rangle + \varepsilon_k \\
&\leq f(x) - \frac{1}{2\gamma} \|z_{k+1} - y_k\|^2 - \left\langle \frac{1}{\gamma}(z_{k+1} - y_k), y_k - x \right\rangle + \left\langle \frac{1}{\gamma}f_k, x - z_{k+1} \right\rangle + \varepsilon_k
\end{aligned}$$

where the first inequality uses (4), the second inequality uses the convexity of $g(x)$, the third inequality uses the convexity of $h(x)$ and (25), the final inequality uses $\gamma < \frac{1}{L}$. Let $x = x_k$ and x^* , we have

$$f(z_{k+1}) - f(x_k) \leq -\frac{1}{2\gamma} \|z_{k+1} - y_k\|^2 - \left\langle \frac{1}{\gamma}(z_{k+1} - y_k), y_k - x_k \right\rangle + \left\langle \frac{1}{\gamma}f_k, x_k - z_{k+1} \right\rangle + \varepsilon_k \quad (50)$$

$$f(z_{k+1}) - f(x^*) \leq -\frac{1}{2\gamma} \|z_{k+1} - y_k\|^2 - \left\langle \frac{1}{\gamma}(z_{k+1} - y_k), y_k - x^* \right\rangle + \left\langle \frac{1}{\gamma}f_k, x^* - z_{k+1} \right\rangle + \varepsilon_k \quad (51)$$

Multiplying (50) by $t_k - 1$ and adding (51), we have

$$\begin{aligned}
&t_k f(z_{k+1}) - (t_k - 1)f(x_k) - f(x^*) \\
&\leq -\frac{t_k}{2\gamma} \|z_{k+1} - y_k\|^2 - \left\langle \frac{1}{\gamma}(z_{k+1} - y_k), (t_k - 1)(y_k - x_k) + y_k - x^* \right\rangle \\
&\quad + \left\langle \frac{1}{\gamma}f_k, (t_k - 1)(x_k - z_{k+1}) + x^* - z_{k+1} \right\rangle + t_k \varepsilon_k
\end{aligned} \quad (52)$$

Thus, we have

$$\begin{aligned}
&t_k (f(z_{k+1}) - f(x^*)) - (t_k - 1)(f(x_k) - f(x^*)) \\
&\leq -\frac{t_k}{2\gamma} \|z_{k+1} - y_k\|^2 - \left\langle \frac{1}{\gamma}(z_{k+1} - y_k), (t_k - 1)(y_k - x_k) + y_k - x^* \right\rangle \\
&\quad + \left\langle \frac{1}{\gamma}f_k, (t_k - 1)(x_k - z_{k+1}) + x^* - z_{k+1} \right\rangle + t_k \varepsilon_k
\end{aligned} \quad (53)$$

Multiplying both sides of (53) by t_k and using $t_k^2 - t_k = (t_{k-1})^2$ in Algorithm 2, we have that

$$\begin{aligned}
&t_k^2 (f(z_{k+1}) - f(x^*)) - t_{k-1}^2 (f(x_k) - f(x^*)) \\
&\leq -\frac{t_k^2}{2\gamma} \|z_{k+1} - y_k\|^2 - \left\langle t_k \frac{1}{\gamma}(z_{k+1} - y_k), (t_k - 1)(y_k - x_k) + y_k - x^* \right\rangle
\end{aligned} \quad (54)$$

$$\begin{aligned}
& + \left\langle t_k \frac{1}{\gamma} f_k, (t_k - 1)(x_k - z_{k+1}) + x^* - z_{k+1} \right\rangle + t_k^2 \varepsilon_k \\
= & - \frac{t_k^2}{2\gamma} \|z_{k+1} - y_k\|^2 - \left\langle t_k \frac{1}{\gamma} (z_{k+1} - y_k), t_k y_k - (t_k - 1)x_k - x^* \right\rangle \\
& + \left\langle t_k \frac{1}{\gamma} f_k, (t_k - 1)x_k - t_k z_{k+1} + x^* \right\rangle + t_k^2 \varepsilon_k \\
= & \frac{1}{2\gamma} \left(\|(t_k - 1)x_k - t_k y_k + x^*\|^2 - \|(t_k - 1)x_k - t_k z_{k+1} + x^*\|^2 \right) \\
& + \left\langle t_k \frac{1}{\gamma} f_k, (t_k - 1)x_k - t_k z_{k+1} + x^* \right\rangle + t_k^2 \varepsilon_k
\end{aligned}$$

Define $U_{k+1} = t_k z_{k+1} - (t_k - 1)x_k - x^*$. Because $y_k = x_k + \frac{t_{k-1}}{t_k}(z_k - x_k) + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$, we have that

$$z_k = \frac{t_k}{t_{k-1}} y_k + \frac{1-t_k}{t_{k-1}} x_k + \frac{t_{k-1}-1}{t_{k-1}} x_{k-1} \quad (55)$$

Thus, we have that $U_k = t_{k-1} z_k - (t_{k-1} - 1)x_{k-1} - x^* = t_k y_k - (t_{k-1} - 1)x_k - x^*$. Substitute U_{k+1} and U_k into (54), we have that

$$\begin{aligned}
& t_k^2 (f(z_{k+1}) - f(x^*)) - t_{k-1}^2 (f(x_k) - f(x^*)) \\
\leq & \frac{1}{2\gamma} \left(\|U_k\|^2 - \|U_{k+1}\|^2 \right) - \left\langle t_k \frac{1}{\gamma} f_k, U_{k+1} \right\rangle + t_k^2 \varepsilon_k
\end{aligned} \quad (56)$$

If $f(z_{k+1}) \leq f(v_{k+1})$, we have $x_{k+1} = z_{k+1}$. Thus,

$$\begin{aligned}
& t_k^2 (f(x_{k+1}) - f(x^*)) - t_{k-1}^2 (f(x_k) - f(x^*)) \\
= & t_k^2 (f(z_{k+1}) - f(x^*)) - t_{k-1}^2 (f(x_k) - f(x^*)) \\
\leq & \frac{1}{2\gamma} \left(\|U_k\|^2 - \|U_{k+1}\|^2 \right) - \left\langle t_k \frac{1}{\gamma} f_k, U_{k+1} \right\rangle + t_k^2 \varepsilon_k
\end{aligned} \quad (57)$$

If $f(z_{k+1}) > f(v_{k+1})$, we have $x_{k+1} = v_{k+1}$. Thus,

$$\begin{aligned}
& t_k^2 (f(x_{k+1}) - f(x^*)) - t_{k-1}^2 (f(x_k) - f(x^*)) \\
= & t_k^2 (f(z_{k+1}) - f(x^*)) - t_{k-1}^2 (f(x_k) - f(x^*)) \\
\leq & \frac{1}{2\gamma} \left(\|U_k\|^2 - \|U_{k+1}\|^2 \right) - \left\langle t_k \frac{1}{\gamma} f_k, U_{k+1} \right\rangle + t_k^2 \varepsilon_k
\end{aligned} \quad (58)$$

Combining (57) and (58), we have

$$\begin{aligned}
& t_k^2 (f(x_{k+1}) - f(x^*)) - t_{k-1}^2 (f(x_k) - f(x^*)) \\
\leq & \frac{1}{2\gamma} \left(\|U_k\|^2 - \|U_{k+1}\|^2 \right) - \left\langle t_k \frac{1}{\gamma} f_k, U_{k+1} \right\rangle + t_k^2 \varepsilon_k \\
\leq & \frac{1}{2\gamma} \left(\|U_k\|^2 - \|U_{k+1}\|^2 \right) + t_k \sqrt{2 \frac{1}{\gamma} \varepsilon_k} \|U_{k+1}\| + t_k^2 \varepsilon_k
\end{aligned} \quad (59)$$

By summing the the inequality (59) over $k = 1, \dots, m$, we obtain

$$\begin{aligned}
& t_m^2 (f(x_{k+1}) - f(x^*)) \\
&= t_m^2 (f(x_{k+1}) - f(x^*)) - t_0^2 (f(x_1) - f(x^*)) \\
&\leq \frac{1}{2\gamma} \left(\|U_1\|^2 - \|U_{m+1}\|^2 \right) + \sum_{k=1}^m t_k \sqrt{2\frac{1}{\gamma}\varepsilon_k} \|U_{k+1}\| + \sum_{k=1}^m t_k^2 \varepsilon_k
\end{aligned} \tag{60}$$

According to (60), we have that

$$\|U_{m+1}\|^2 \leq \underbrace{\|U_1\|^2 + 2\gamma \sum_{k=1}^m t_k^2 \varepsilon_k}_A + \sum_{k=1}^m \underbrace{2\gamma t_k \sqrt{2\frac{1}{\gamma}\varepsilon_k} \|U_{k+1}\|}_{\lambda_k} \tag{61}$$

According to Lemma 1 in (Schmidt et al., 2011), we have that

$$\begin{aligned}
& \|U_{m+1}\| \\
&\leq \frac{1}{2} \sum_{k=1}^m \lambda_k + \left(A + \left(\frac{1}{2} \sum_{k=1}^m \lambda_k \right)^2 \right)^{\frac{1}{2}} \\
&= \frac{1}{2} \sum_{k=1}^m \underbrace{2\gamma t_k \sqrt{2\frac{1}{\gamma}\varepsilon_k}}_{A_m} + \left(\|U_1\|^2 + \underbrace{2\gamma \sum_{k=1}^m t_k^2 \varepsilon_k}_{B_m} + \left(\underbrace{\frac{1}{2} \sum_{k=1}^m 2\gamma t_k \sqrt{2\frac{1}{\gamma}\varepsilon_k}}_{A_m} \right)^2 \right)^{\frac{1}{2}} \\
&\leq A_m + \left(\|U_1\|^2 + B_m + A_m^2 \right)^{\frac{1}{2}}
\end{aligned} \tag{62}$$

Because A_k and B_k are increasing sequences, $\forall k \leq m$, we have that

$$\begin{aligned}
\|U_k\| &\leq A_m + \left(\|U_1\|^2 + B_m + A_m^2 \right)^{\frac{1}{2}} \leq A_m + \|U_1\| + \sqrt{B_m} + A_m \\
&\leq 2A_m + \|U_1\| + \sqrt{B_m}
\end{aligned} \tag{63}$$

According to (60) and (63), we have that

$$\begin{aligned}
& t_m^2 (f(x_{k+1}) - f(x^*)) \\
&\leq \frac{1}{2\gamma} \left(\|U_1\|^2 - \|U_{m+1}\|^2 \right) + \sum_{k=1}^m t_k^2 \varepsilon_k + \sum_{k=1}^m t_k \sqrt{2\frac{1}{\gamma}\varepsilon_k} \|U_{k+1}\| \\
&\leq \frac{1}{2\gamma} \|U_1\|^2 + \frac{1}{2\gamma} B_m + \frac{1}{\gamma} A_m \left(2A_m + \|U_1\| + \sqrt{B_m} \right) \\
&\leq \frac{1}{2\gamma} \left(2A_m + \|U_1\| + \sqrt{B_m} \right)^2
\end{aligned} \tag{64}$$

Because $t_{k+1} = \frac{\sqrt{4t_k^2+1}+1}{2}$, it is easy to verify that $t_k \geq \frac{k+1}{2}$. Thus, we have

$$f(x_{k+1}) - f(x^*) \leq \frac{2L}{(m+1)^2} \left(\|x_0 - x^*\| + 2A_m + \sqrt{B_m} \right)^2 \tag{65}$$

This completes the proof. ■

5. Conclusion

In a lot of problems the proximal operator does not have an analytic solution, or is expensive to obtain an exact solution. In this paper, we propose inexact proximal gradient methods (not only a basic inexact proximal gradient method (IPG), but also a Nesterov’s accelerated inexact proximal gradient method (AIPG)) for non-convex and non-smooth optimization, which tolerate an error in the calculation of the proximal operator. Theoretical analysis shows that IPG and AIPG have the same convergence rates as in the error-free case, provided that the errors decrease at appropriate rates.

References

- Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1035–1043, 2011.
- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
- Dimitri P Bertsekas, Angelia Nedi, Asuman E Ozdaglar, et al. Convex analysis and optimization. 2003.
- Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2): 459–494, 2014.
- Radu Ioan Boţ, Ernő Robert Csetnek, and Szilárd Csaba László. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4(1):3–25, 2016.
- Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *AISTATS*, pages 57–64, 2005.
- Olivier Chapelle, Mingmin Chi, and Alexander Zien. A continuation method for semi-supervised svms. In *Proceedings of the 23rd international conference on Machine learning*, pages 185–192. ACM, 2006.
- Hong Chen and Yulong Wang. Kernel-based sparse regression with the correntropy-induced loss. *Applied and Computational Harmonic Analysis*, 2016.
- John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.
- John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

- Yunlong Feng, Xiaolin Huang, Lei Shi, Yuning Yang, and Johan AK Suykens. Learning with the maximum correntropy criterion induced losses for regression. *Journal of Machine Learning Research*, 16:993–1034, 2015.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.
- Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.
- Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1561–1576, 2011.
- Cho-Jui Hsieh and Peder A Olsen. Nuclear norm minimization via active subspace selection. In *ICML*, pages 575–583, 2014.
- Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.
- Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2012.
- Rodolphe Jenatton, Julien Mairal, Francis R Bach, and Guillaume R Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 487–494, 2010.
- Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in Neural Information Processing Systems*, pages 379–387, 2015.
- Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pages 1–40, 2011.
- Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *arXiv preprint arXiv:1109.2415*, 2011.
- Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact coordinate descent: complexity and preconditioning. *Journal of Optimization Theory and Applications*, pages 144–176, 2016.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- Silvia Villa, Saverio Salzo, Luca Baldassarre, and Alessandro Verri. Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23(3):1607–1633, 2013.
- Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(Mar):1081–1107, 2010.

Leon Wenliang Zhong and James T Kwok. Efficient sparse modeling with automatic feature grouping. *IEEE transactions on neural networks and learning systems*, 23(9):1436–1447, 2012.